

Journal of Experimental Psychology: General

Metacognitive Improvement: Disentangling Adaptive Training From Experimental Confounds

Martin Rouy, Vincent de Gardelle, Gabriel Reyes, Jérôme Sackur, Jean Christophe Vergnaud, Elisa Filevich, and Nathan Faivre

Online First Publication, February 14, 2022. <http://dx.doi.org/10.1037/xge0001185>

CITATION

Rouy, M., de Gardelle, V., Reyes, G., Sackur, J., Vergnaud, J. C., Filevich, E., & Faivre, N. (2022, February 14). Metacognitive Improvement: Disentangling Adaptive Training From Experimental Confounds. *Journal of Experimental Psychology: General*. Advance online publication. <http://dx.doi.org/10.1037/xge0001185>

Metacognitive Improvement: Disentangling Adaptive Training From Experimental Confounds

Martin Rouy¹, Vincent de Gardelle², Gabriel Reyes³, Jérôme Sackur⁴, Jean Christophe Vergnaud⁵,
Elisa Filevich^{6, 7}, and Nathan Faivre¹

¹ Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LPNC

² Paris School of Economics and CNRS

³ Faculty of Psychology, Universidad Del Desarrollo

⁴ Laboratoire de Sciences Cognitives et Psycholinguistique, École Normale Supérieure, PSL University

⁵ Centre d'Economie de la Sorbonne, Paris, France

⁶ Department of Psychology, Humboldt Universität zu Berlin

⁷ Bernstein Center for Computational Neuroscience Berlin, Berlin, Germany

Metacognition is defined as the capacity to monitor and control one's own cognitive processes. Recently, Carpenter and colleagues (2019) reported that metacognitive performance can be improved through adaptive training: healthy participants performed a perceptual discrimination task, and subsequently indicated confidence in their response. Metacognitive performance, defined as how much information these confidence judgments contain about the accuracy of perceptual decisions, was found to increase in a group of participants receiving monetary reward based on their confidence judgments over hundreds of trials and multiple sessions. By contrast, in a control group where only perceptual performance was incentivized, metacognitive performance remained constant across experimental sessions. We identified two possible confounds that may have led to an artificial increase in metacognitive performance, namely the absence of reward in the initial session and an inconsistency between the reward scheme and the instructions about the confidence scale. We thus conducted a preregistered conceptual replication where all sessions were rewarded and where instructions were consistent with the reward scheme. Critically, once these two confounds were corrected we found moderate evidence for an absence of metacognitive training. Our data thus suggest that previous claims about metacognitive training are premature, and calls for more research on how to train individuals to monitor their own performance.

Keywords: cognitive training, confidence, introspection, metacognition

Metacognition is defined as the capacity to monitor and control one's own cognitive processes (Flavell, 1979; Nelson & Narens, 1994). Metacognitive monitoring is imperfect: Under- or overestimations regarding the accuracy of one's own judgments are frequent, both in healthy individuals (Shekhar & Rahnev, 2021a, 2021b) and in individuals with neurological or

psychiatric disorders (Hoven et al., 2019; Rouy, Saliou, et al., 2021). Thus, one outstanding issue is whether one can design training protocols to help individuals improve their abilities to evaluate their own performances.

Recently, Carpenter and colleagues (2019) proposed that metacognitive abilities can be improved through adaptive training. In

Martin Rouy  <https://orcid.org/0000-0003-4280-4683>

Nathan Faivre  <https://orcid.org/0000-0001-6011-4921>

Elisa Filevich and Nathan Faivre contributed equally.

All authors developed the study concept and contributed to the study design. Modifications in the original code were implemented by Elisa Filevich. Data collection was performed by Martin Rouy. Martin Rouy and Nathan Faivre analyzed data. Martin Rouy and Nathan Faivre drafted the article; all authors provided critical revisions and approved the final version of the article for submission. The authors declare no competing interests.

Nathan Faivre has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant 803122). Elisa Filevich was supported by a Freigeist

Fellowship from the Volkswagen Foundation (Grant 91620). Jérôme Sackur received support from the Agence Nationale de la Recherche, ANR-17-EURE-0017. We thank Steve Fleming for sharing the materials of the original study and commenting on a first version of this article.

This work has been presented at the occasion of the 24th annual meeting of the Association for the Scientific Study of Consciousness.

Preregistration is publicly available: <https://osf.io/gak2t>.

Data and analysis scripts are publicly available: <https://doi.org/10.17605/OSF.IO/RQ967> (Rouy, de Gardelle, et al., 2021).

Correspondence concerning this article should be addressed to Martin Rouy, Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LPNC UMR 5105 UGA BSHM, 1251 Avenue Centrale, 38058 Grenoble Cedex 9, France. Email: martinrouy03@gmail.com

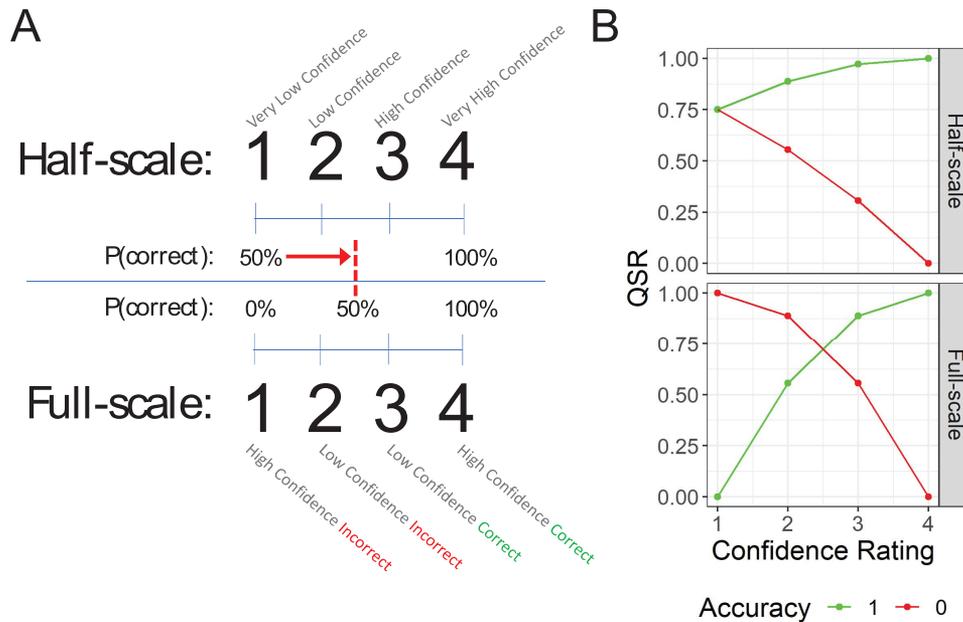
their study, healthy participants were asked to perform both a memory and a perceptual discrimination task, either with shapes or words stimuli, and subsequently report their confidence in their response. They used a longitudinal protocol in 10 sessions (see Figure 2A), where the first session (S1, or pretraining session) served as a baseline, followed by eight sessions of training (S2–S9) on the perceptual task, and finally a posttraining session (S10). In the training sessions, participants received feedback and monetary reward on the basis of their confidence evaluations, after each block of 27 trials: the better the confidence ratings reflected perceptual accuracy in that block, the higher the reward. The pretraining and posttraining sessions had no feedback.

Importantly, Carpenter and colleagues reported that metacognitive efficiency, defined as the adequacy between task performance and confidence, increased between pre- and posttraining sessions in the experimental group where participants received monetary reward on their metacognitive performance, but remained constant in a control group rewarded on their perceptual performance.

In their article, Carpenter et al. argued that the increase in metacognitive efficiency that they observed in the posttraining session (S10) was mediated by an increase in overall confidence between the pretraining session (S1) and the following session (S2) occurring only in the experimental group. A close inspection of these results reveal that confidence indeed sharply increased from the very beginning of S2, and remained constant afterward. Likewise, metacognitive performance increased between the pretraining session and S2 but remained constant from S2 onward. This sudden increase in confidence and metacognitive performance suggests that they might have occurred due to factors other than training.

We identified two potential confounding factors which we thought could lead to apparent increases in metacognitive efficiency, without involving a real improvement as a result of training. First, because no reward was offered during the pretraining session, it is possible that the sharp increase in average confidence in S2 reflects a response bias due to the introduction of incentives. Indeed, recent research shows that positive (resp. negative) reward increased (resp. decreased) confidence irrespective of task performance or metacognitive abilities (Lebreton et al., 2018). Second, the increase in confidence may be driven by differences in the definition of the possible confidence ratings across groups. Indeed, in the pretraining session participants in both the experimental and control groups were instructed to report confidence on a four-level scale, defined as 1 = *very low confidence*, 2 = *low confidence*, 3 = *high confidence* and 4 = *very high confidence*. Importantly no explicit mapping from confidence levels to subjective probabilities was given to participants. In this context, the correct interpretation of the lowest confidence rating is that of a 50% chance of being correct, that is, being unsure of the accuracy of their response, and therefore that participants are provided with a half-scale of confidence (Figure 1A). Yet, from S2 to S9, the experimental group (but not the control group) was presented with a full confidence scale, that is, confidence was mapped onto a probability of a response being correct from 0 to 1. As a result, confidence ratings 1 and 2 were to be used in case subjects thought they made an error (level 1 would be used when they were certain that they made an error, see Figure 1A), which rarely occurs in such experimental settings. This full-scale was explained to participants at the beginning of S2 and implemented in the reward scheme. For instance, according to a full-scale, rating confidence 1 (i.e., “sure

Figure 1
Confidence Rating Scales



Note. (A) Meaning of each confidence rating depending on the type of confidence scale (Half vs. Full), along with the corresponding probability of being correct (P(correct)). (B) Reward schemes depending on the type of confidence scale (Half vs. Full). QSR = Quadratic Scoring rule. See the online article for the color version of this figure.

incorrect”) when incorrect is maximally rewarded ($QSR = 1$, see Method) while rating confidence 1 on a half-scale (i.e., “not sure”) is equally rewarded regardless of accuracy (Figure 1B). Using a full-scale, participants should mostly use the highest ratings, as one can assume that the confident detection of errors is rare in nonspeeded perceptual tasks. Thus, ratings should increase from the first to the second session.

Thus, the introduction of incentives and the switch from a half-scale to a full-scale may have led to an artificial increase in confidence bias. Importantly, this upward shift in confidence ratings may also be expected to produce an artificial increase in metacognitive efficiency. Indeed, precise confidence criteria might be easier to maintain across two levels on a full scale than four levels on a half-scale. In addition, as suggested in recent works (Shekhar & Rahnev, 2021a, 2021b; Xue et al., 2021) criteria for high confidence are noisier than criteria for low confidence and thus a merge of high confidence categories can artificially increase metacognitive efficiency.

To assess the contribution of these potential confounds to the observed effects, we reanalyzed the original data, and collected a new dataset attempting to replicate the original findings while controlling for both keeping the incentives and reward scheme constant across sessions (Figure 2B). Assuming that the original procedure involves genuine metacognitive training, we reasoned that metacognitive efficiency should increase between the first and last session in the experimental group even when issues related to incentives and reward scheme are corrected. Instead, based on a preregistered sample size of 18 participants, we provide moderate evidence in favor of the null hypothesis according to which adaptive training in the present form does not improve metacognitive ability.

Method

Metacognitive Performance Measurement

To evaluate metacognitive performance, we relied on the M-Ratio measure, derived from the *meta-d'* framework by Maniscalco and Lau (2012). In signal detection theory, the sensitivity d' quantifies the ability to detect or discriminate a stimulus from the distributions of correct and incorrect responses. Likewise, the metacognitive sensitivity *meta-d'*, quantifies the expected discriminability between two stimuli, if sensory evidence were not degraded between the discrimination decision and confidence rating. Thus, *meta-d'* refers to the sensory evidence available for metacognition, just as d' is the sensory evidence available for decision-making. It is then possible to quantify how much information was available for the metacognitive task, relative to the information available for the type I task, using the ratio *meta-d'*/ d' . This measure, called M-Ratio, is considered as the efficiency of metacognition for each observer.

Reanalysis of Original Data

We retrieved the original data from the authors and further characterized the evolution of metacognitive performance across sessions with additional mixed-model ANOVAs with Training (Pretraining session vs Posttraining session) and group (Control vs Experimental) as factors. In line with the original mediation analysis, we expected to find a significant increase in metacognitive performance between pretraining session and S2. Furthermore, we compared S2 and S9 to assess the effect of training itself irrespective of the difference in incentives between pretraining session and

Figure 2
Comparison of the Original Study by Carpenter et al and the Present Study

A. Carpenter et al.



B. Present study



Note. (A) Original version of the protocol, with pre- and posttraining sessions providing no feedback, and rewards from S2 to S9 mapped onto a full-confidence scale. (B) Present version of the protocol, with S1 and S10 providing feedback, and rewards from S2 to S9 mapped onto a half-confidence scale. From “Domain-general enhancements of metacognitive ability through adaptive training”, by Carpenter, J., Sherman, M. T., Kievit, R. A., Seth, A. K., Lau, H., & Fleming, S. M., 2019, *Journal of Experimental Psychology: General*, 148(1), 51. Copyright 2019 by the authors. Adapted with permission. See the online article for the color version of this figure.

S2. Statistical analyses were conducted on $\log(\text{meta-}d'/d')$, like in the original study.

Conceptual Replication

Methods and hypotheses were preregistered (<https://osf.io/gak2t>) prior to data collection.

Modifications From the Original study

First, to test the possibility that a difference in terms of incentives between the pre and posttraining sessions might have artificially inflated metacognitive performance, we kept the incentives constant throughout the 10 sessions of the experiment. Accordingly, we refer to the first and last sessions as S1 and S10, instead of the original “pretraining” and “posttraining” sessions, respectively (see Figure 2). In the pre- and posttraining sessions, participants in the original study could either start with the memory tasks or the perception tasks. As a consequence of rewarding S1 and S10, participants always started with the perception task. This is to allow for continuity in the explanation of how points were calculated and assigned to participants.

Detailed instructions on how to map confidence to correct and incorrect trials were provided after the titration tasks in S1 but before any task where participants rated confidence. As in the original study, these instructions included a predefined set of demonstration trials and a series of practice trials with trial-wise feedback about whether confidence ratings were correctly assigned to correct or incorrect trials. However, here we made sure that the instructions were consistent with the reward scheme, and that both corresponded to a half-scale.

Second, to assess whether the increase in metaperformance observed in the original study stemmed from an incongruence between instructions regarding the confidence scale and reward, we provided reward that was consistent with instructions in all sessions: Participants were instructed to report confidence on a four-point scale with 1 = *very low confidence*, 2 = *low confidence*, 3 = *high confidence* and 4 = *very high confidence*, in all sessions including S1 and S10 (see Figure 2B). As opposed to the original study, we mapped confidence onto a probability of being correct between .5 and 1, as follows: $P(\text{correct}) = \frac{\text{conf} + 2}{6}$. Subsequently the quadratic scoring rule (QSR) was defined as $1 - (\text{accuracy} - P(\text{correct}))^2$, for each trial (see Figure 1B).

We also performed minor modifications to the experiment with no consequence on the experimental design: for example, Carpenter and colleagues ran the initial titration staircase until a fixed number of reversals was reached, or a maximum of 60 trials. We ran the titration staircase for a fixed number of 60 trials. We also fixed a small error in the code shared by Carpenter and colleagues in the memory task resulting in images being presented more than once in each block, and other images to never be displayed. All corresponding details are provided in our preregistration document (<https://osf.io/gak2t>).

Participants

The sample size was determined according to a preregistered stopping rule, using an open-ended sequential Bayes Factor (BF) design. Thus, we tested our effect of interest, namely the interaction between groups (Control vs. Experimental) and sessions (S1

vs. S10) on metacognitive efficiency until moderate evidence toward H1 or H0 was reached, that is, $\text{BF} > 5$ or $\text{BF} < .2$, respectively. As in the original study, we recruited participants through Amazon’s MTurk participant marketplace. Sixty-nine participants completed at least the first session. Of these, 11 participants dropped out from the study before the end of the tenth session. Nine participants were excluded for responding incorrectly to screening questions related to the understanding of the tasks, before the beginning of the training (for details, see Carpenter et al., 2019). Nineteen participants were excluded for technical issues during the first session, leading them to drop at least one experimental condition. Further, 11 participants were excluded for either floor ($< 55\%$) or ceiling ($> 95\%$) performance in at least one condition/session. Finally, one participant was excluded for reporting the same confidence level on at least 95% of the trials over three sessions or more. Trials where participants did not respond in time ($> 2,000$ ms) or responded too quickly (< 200 ms) were excluded from further analyses (1.61% of the trials).

The analyses were conducted on a sample of 18 participants (10 women, mean age = 40.4 years, range age = 19–59). All participants received monetary compensation in U.S. dollars (range = \$37.6–\$41.8). An upper bound for sample size was determined using a design analysis with Bayes factors as index of evidence (Schönbrodt & Wagenmakers, 2018). Data simulations with an expected increase in metacognitive efficiency between S1 and S10 of small effect size (Cohen’s $d = .3$) revealed that a maximal sample of 100 participants would lead to conclusive evidence under H1 in 74% of cases ($\text{BF} > 5$), and under H0 in 89% of cases ($\text{BF} < .2$). However, the stopping rule criterion was already met when performing the first Bayes Factor sequential analysis after a first group of 18 participants had completed all ten sessions (see Figure 5). We recruited participants in the experimental group only (i.e., participants receiving reward according to metacognitive performance), and compared their data with those of participants in the original control group, who received reward according to their perceptual performance. As in the original study, bonuses were distributed pseudorandomly to ensure equivalent financial motivation irrespective of performance. The study was approved by the ethics committee from the Paris School of Economics (#2019 021).

Procedure

Save from the modifications to the code, we used the same HTML/JS/CSS scripts, and therefore the very same stimuli, as in the original study by Carpenter et al. The study ran on a JATOS server (www.jatos.org; Lange et al., 2015).

Statistical Analysis

We ran the same analyses as Carpenter and colleagues. We tested for potential changes in metacognitive efficiency ($\log(\text{meta-}d'/d')$) and metacognitive bias (average confidence) using mixed-design ANOVAs in Rstudio Version 1.3.1093 (RStudio Team, 2020) using notably the packages tidyverse (Wickham et al., 2019), afex (Singmann et al., 2015), and metaSDT (Craddock, 2018). Bayesian ANOVAs were computed with default prior (Cauchy distribution centered on the effect size, with a scaling parameter set to $\frac{\sqrt{2}}{2}$) using the BayesFactor package (Morey et al., 2018).

Results

Reanalysis of Carpenter et al. (2019)

After confirming the results reported by Carpenter et al. we extended the analyses reported in the original paper in two ways. First, to account for a potential effect of a change in instructions in S2 versus Pretraining, we compared metacognitive efficiency between S2 and the posttraining session S10 (instead of between pre- and posttraining sessions, as originally reported). Here, we found no significant interaction effect between group and Training, $F(1, 58) = .71, p = .40, BF = .27$. When comparing S2 with S9 (i.e., the first and the last of the training sessions), the Group \times Training interaction remained nonsignificant, $F(1, 59) = .49, p = .49, BF = .39$ (Figure 3A and 3B). These results suggest that the improvement of metacognitive efficiency occurred not during the extended training part of the protocol, but quite abruptly at the beginning of the training phase.

Second, we studied the abrupt changes in metacognitive efficiency between the pretraining session and S2. We first found a significant interaction between group and Training, $F(1, 59) = 4.64, p = .035$. Perhaps more strikingly, we found in the original data an abrupt increase in average confidence between the last five trials of the pretraining session and the first five trials of S2 (Figure 4E), in the experimental group only, $F(1, 28) = 22.14, p < .001$. Together, these results suggest that this increase in metacognitive efficiency could be driven by the changes introduced from S2 to S9, also influencing participants' strategy on the posttraining session (S10).

A Preregistered Replication Study

Sequential Bayes Factor Analysis

Informed by the reanalysis of the original data, we then turned to our conceptual replication study. To assess the efficiency of metacognitive training while accounting for incentives and confidence scale confounding factors, we conducted the same analysis

as in the original study comparing metacognitive efficiency ($\log(\text{meta-}d'/d')$) between sessions (S1 and S10) and groups (experimental vs. control).

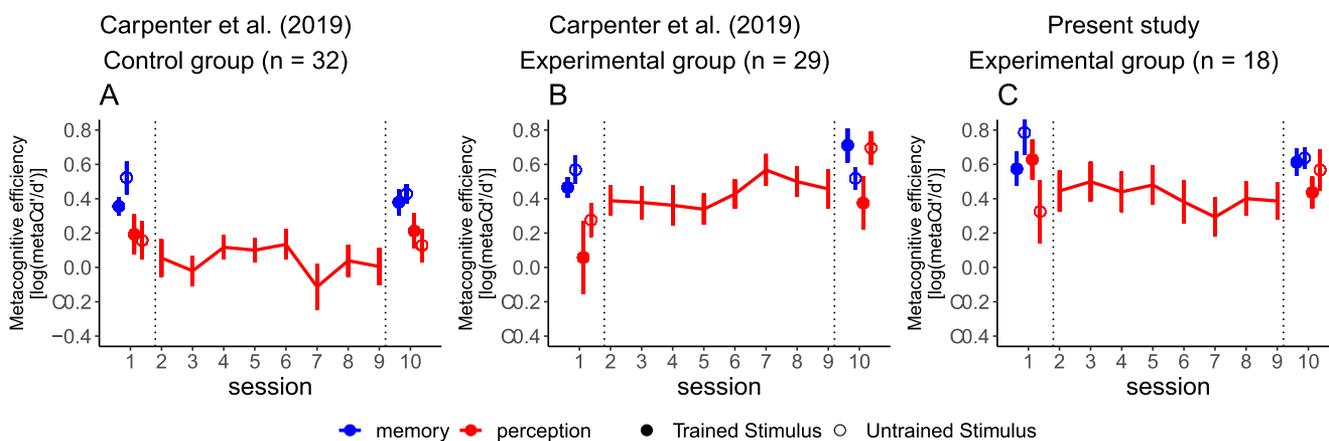
We had preregistered recruiting participants until moderate evidence toward H1 or H0 was reached.

Metacognitive Efficiency

We compared metacognitive efficiency in S1 and S10 in our new experimental group (Figure 3C) with those in the control group from Carpenter et al. (2019). (Figure 3A). Contrary to the original results, the group \times Training interaction was not significant in this analysis, $F(1, 45) = .083, p = .93, BF = .17$. Moreover, assessing the linear trend of metacognitive efficiency between S2 and S9 in the three groups, we found no main effect of the training sessions, $F(7, 490) = .25, p = .97, BF = .13$, and no interaction effect between the training sessions and groups (control vs. experimental group in the original study: $F[7, 399] = 1.61, p = .13, BF = 2.50$; control vs. our experimental group: $F[7, 294] = .90, p = .51, BF = .24$). In other words, once we kept the reward scheme constant across all sessions, we found no evidence for metacognitive training in our study. This suggests that previous results might have been confounded by effects of incentives and/or confidence scale, as we detailed in the Introduction.

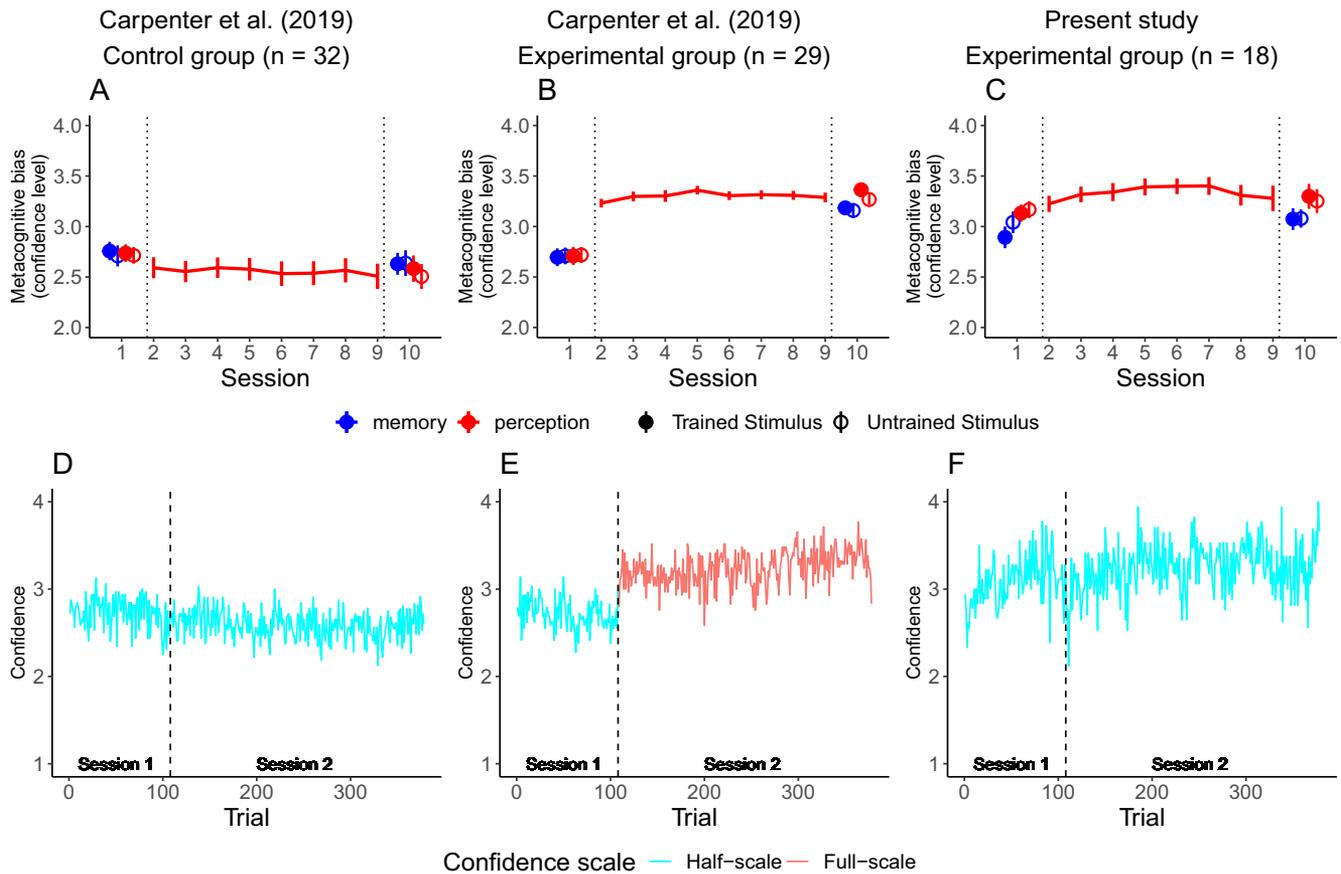
In their study, Carpenter and colleagues also reported that the peak change in metacognitive efficiency occurred systematically later than the peak change in confidence bias. To assess if a similar pattern was present in our replication group, we conducted an ANOVA with peak session as dependent variable, and outcome (metacognitive efficiency vs. confidence bias) and group (experimental: original vs. replication) as fixed effects. This analysis revealed a main effect of outcome, $F(1, 45) = 11.37, p = .02$, but no interaction with group, $F(1, 45) = .01, p = .98$, indicating that in both groups the peak change in metacognitive efficiency occurred systematically later than the peak change in confidence bias. Because this temporal pattern was also found in our replication group in the absence of global increase in metacognitive efficiency, the extent to which those dynamics are important for

Figure 3
Metacognitive Efficiency ($\log(\text{meta-}d'/d')$) Over the Ten Experimental Sessions



Note. (A and B) Results reproduced from the original data by Carpenter et al, control group, and experimental group, respectively. (C) Results from the present study. Error bars represent standard error of the mean. See the online article for the color version of this figure.

Figure 4
Confidence Level Across Sessions and Trials



Note. A–C: Metacognitive bias across sessions in the control (A) and experimental groups (B) from Carpenter et al., and in the experimental group from our sample (C). Note. Error bars represent standard error of the mean. D–F: Evolution of average confidence across participants and trials in S1 and S2 in the control (D) and experimental groups (E) from Carpenter et al., and in the experimental group from our sample (F). Colors indicate the type of confidence scale in use. Blue: Half-scale, Red: Full-scale. See the online article for the color version of this figure.

metacognitive training remains unclear. Of note, these results are based on a rather small sample size, in compliance with the stopping rule we preregistered prior to data collection.

Exploring the Origin of the Metacognitive Bias

Next, we assessed in an exploratory analysis which of the two confounds, incentives or confidence scale, was the main contributor of the confidence increase. This also relates to the question of metacognitive training, as Carpenter and colleagues reported that the increase in metacognitive efficiency was in fact mediated by the increase in metacognitive bias, and as an increase in confidence bias might result in an increase in metacognitive efficiency (Shekhar & Rahnev, 2021).

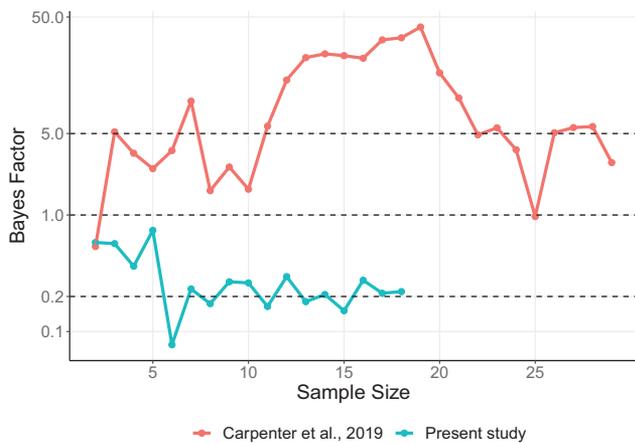
If this abrupt increase in confidence ratings was due to the introduction of incentives at S2, we would expect the same average confidence in our experimental group (Figure 4C) and from S2 to S9 in the original experimental group, as these conditions are similar in terms of reward. We would also expect these two conditions to show higher levels of confidence than the control group. This is what we found in the data. When comparing average confidence in

S2–S9 between the three groups (control vs. original experimental vs. replication) with an ANOVA, we found a main effect of group, $F(2, 72) = 24.61, p < .001$, driven by significantly higher levels of confidence both in our replication group, $t(72) = -5.05, p < .001$, and in the original experimental group, $t(72) = -6.43, p < .001$, compared with the control group, with no difference between the experimental group and the replication group, $t(72) = .10, p = .995, BF = .46$. However, we are cautious in interpreting confidence biases that might not be comparable between groups and studies.

One other possibility is that this abrupt increase in average confidence was due to a shift in the type of confidence scale (i.e., half-scale in the pretraining session, and full-scale from S2 to S9, see Figure 2A). If this were true, then we would expect the average confidence in our replication group (which used a half-confidence scale) to be lower than the level of confidence obtained from S2 to S9 in the original experimental group. As just mentioned, however, these two conditions were not different in terms of average confidence.

Furthermore, because the increased levels of confidence described above are not accompanied by an increase in first-order

Figure 5
Bayes Factor (BF) Sequential Analysis of the Interaction Effect Between Sessions (S10 and S1) and Groups (Control Versus Experimental) on $\log(\text{meta-}d'/d)$



Note. The BF assesses whether the effect of interest (interaction Group \times Training for metacognitive efficiency) is more plausible under H0 or under H1. $BF > 1$ is evidence supporting H1. $0 < BF < 1$ is evidence supporting H0. The dashed lines mark the ratios where the evidence is five-fold more likely under each hypothesis, which we took as boundaries for moderate evidence. Red curve: Carpenter et al., 2019. Blue curve: Present study. See the online article for the color version of this figure.

performance (as assessed through difficulty levels across the three groups, $F(2, 72) = .17, p = .84, BF = .18$) it is unlikely that the metacognitive bias can simply be explained by a generic motivation effect.

Altogether, these analyses thus suggest that the presence of incentives might be the main reason for the increase in confidence ratings, which in turn would have led to an increase of metacognitive efficiency, as recently proposed (Shekhar & Rahnev, 2021a). Nonetheless, because our analyses relied on comparing confidence biases between studies in relatively small samples, these conclusions on the specific mechanism at stake should be taken with caution.

Discussion

In the present work, we aimed at reassessing the effectiveness of a protocol designed by Carpenter and colleagues (2019) to improve metacognitive abilities. We noticed that the increase in metacognitive efficiency found by Carpenter and colleagues might be unspecific, owing to an artificial increase in confidence bias, triggered by two confounding factors: In the original study, reward was not held constant throughout all sessions, so that participants might have been more incentivized to perform the task not only during rewarded sessions (S2–S9), but also in the posttraining session (S10), as a spillover effect. Also, the instructions provided to the participants in the experimental group were not congruent with the reward scheme, encouraging them to use high confidence ratings (i.e., ratings 3 and 4) from S2 onward but not in the pretraining session. To evaluate our claim that the original results may be due to confounding factors, we performed additional analyses on the original data set. First, when restricting the analysis to training

sessions only (i.e., S2 to S9, instead of pretraining and posttraining sessions), thus controlling for incentives, we found no evidence for an improvement in metacognitive performance in the experimental group. By contrast, this increase was already significant between S1 and S2. This sharp increase in metacognitive performance was accompanied by an abrupt increase in average confidence between the last trials of the pretraining session and the first trials of S2. In our view, the fact these behavioral changes occurred rapidly in time at the very beginning of the experimental procedure casts doubts on the possibility that they arose as a result of a genuine improvement in metacognitive performance. Instead, we suspect that they may have been attributable to either, or both, of the two possible experimental confounds mentioned above.

To further assess the validity of this training procedure, we conducted a conceptual replication controlling for both incentives and confidence-related factors by, first, providing reward in all sessions (i.e., including S1 and S10) and, second, rewarding the experimental group on the basis of a half-confidence scale, in line with the instructions received by participants (and instead of a full-scale as in the original study). We reasoned that, if the training method was effective in improving metacognition, estimates of metacognitive efficiency should increase between S1 and S10 in the experimental group, even when issues related to incentives and confidence scale were corrected. Instead, we obtained moderate evidence in favor of H0 (following a preregistered open-ended sequential Bayes factor analysis), indicating that no increase in metacognitive efficiency occurred. Thus, we suggest that the increase in metacognitive efficiency reported by Carpenter et al. (2019) resulted from a global change in the use of the confidence scale, possibly owing to incentives or instructions regarding the confidence scale, rather than from an improved sensitivity to trial-wise fluctuations in the quality of the decision. While such a global adjustment of confidence ratings might be adaptive and useful (e.g., when communicating confidence to reach joint decisions), it is important to distinguish this effect from a genuine improvement of metacognitive monitoring, conceptually and empirically. Of note, post hoc analyses revealed that metacognitive efficiency in S1 was higher in the replication compared with the original experimental group with marginal significance ($p = .11$), probably attributable to the fact that S1 in our replication group was rewarded, pushing participants to perform better. Yet, it might be that metacognitive efficiency in the replication group reached a ceiling early in the procedure, leaving little room for improvement even if training were in fact possible under this new protocol.

In recent years, the field of metacognition has seen a dramatic increase in popularity, in part due to the development of new statistical tools that allow quantifying metacognitive performance independently from typical confounds such as first-order performance (Fleming & Lau, 2014; Galvin et al., 2003; Maniscalco & Lau, 2012). Moreover, metacognitive deficits are prevalent in several psychiatric and neurological disorders, with severe consequences in terms of medical observance and quality of life (Hasson-Ohayon et al., 2015; Lysaker et al., 2015). This is why developing robust, efficient, and cost-effective remediation procedures to improve metacognitive performance is important. Several studies already provided evidence suggesting that monitoring abilities can be trained: A two-week meditation training was found to enhance metacognitive accuracy in the memory domain (Baird et al., 2014), and knowledge about cognitive biases is held to reduce

delusions and positive symptoms in schizophrenia (for a review, see Eichner & Berna, 2016). More recently, preliminary results from a virtual-reality assisted training consisting in frequently questioning the reality of wakeful experiences augmented the rate of lucid dreaming experiences (Gott et al., 2021). Despite pioneering experiments showing promising results (Adams & Adams, 1958; Sharp et al., 1988), to our knowledge, no recent remediation procedure based on feedback has been successful in improving the quality of confidence ratings (for a recent attempt based on single-trial feedback, see Haddara and Rahnev, 2019, 2020).

Future attempts to improve the quality of confidence ratings may be informed by recent findings regarding the definition of metacognitive noise (Shekhar & Rahnev, 2021a, 2021b; Xue et al., 2021), as a way to provide more information to participants regarding the qualitative nature of their metacognitive deficits. They could also rely on elicitation methods that encourage participants to report optimal confidence estimates, such as measuring participants' willingness to trade a gamble based on the accuracy of their response against a lottery with known probabilities (Dienes & Seth, 2010; Massoni et al., 2014). Another way of refining confidence ratings may be to provide participants with feedback regarding the temporal dynamics with which first-order decisions are made. Indeed, becoming aware of how the decision-making process unfolds in time may help to better judge the accuracy of a given decision. Practically, this could simply consist in presenting participants with feedback about their own response times for correct and incorrect responses, or more ambitiously with parameter estimates from mouse-tracking (Dotan et al., 2019; Faivre et al., 2021) or postdecisional evidence accumulation models (Pleskac & Busemeyer, 2010; Pereira et al., 2020, 2021). Other strategies may consist in training participants to better detect their attentional lapses (Baird et al., 2014; Recht et al., 2021), or to regulate brain networks associated with over or underconfidence (Cortese et al., 2016). Given the complexity of this endeavor, and the societal and clinical issues at stake, effective metacognitive training will probably require collective efforts rather than individual initiatives (Rahnev et al., 2021). In that regard, we highlight the openness from the authors of the original study, who publicly shared their valuable code and data and discussed these results openly with us, as those are the first necessary steps toward collective research on metacognition.

Context of the Research

We were interested in the possibility to train metacognitive abilities in the broader context of our research on schizophrenia. A rich clinical literature suggests the existence of metacognitive deficits in individuals with schizophrenia, and efforts had already been made to alleviate symptoms and improve quality of life through metacognitive training. Existing metacognitive training procedures rely on explicit and high-level strategies, notably by encouraging patients to bring unnoticed beliefs and cognitive biases to awareness. As a complementary intervention, we were enthusiastic about the metacognitive training proposed by Carpenter and colleagues, which targeted lower-level mechanisms involved in learning how to properly estimate confidence on a trial-to-trial basis. If successful in healthy participants, we were hoping to adapt this procedure to clinical settings.

References

- Adams, P. A., & Adams, J. K. (1958). Training in confidence-judgments. *The American Journal of Psychology*, *71*(4), 747–751. <https://doi.org/10.2307/1420334>
- Baird, B., Mrazek, M. D., Phillips, D. T., & Schooler, J. W. (2014). Domain-specific enhancement of metacognitive ability following meditation training. *Journal of Experimental Psychology: General*, *143*(5), 1972–1979. <https://doi.org/10.1037/a0036882>
- Carpenter, J., Sherman, M. T., Kievit, R. A., Seth, A. K., Lau, H., & Fleming, S. M. (2019). Domain-general enhancements of metacognitive ability through adaptive training. *Journal of Experimental Psychology: General*, *148*(1), 51–64. <https://doi.org/10.1037/xge0000505>
- Cortese, A., Amano, K., Koizumi, A., Kawato, M., & Lau, H. (2016). Multivoxel neurofeedback selectively modulates confidence without changing perceptual performance. *Nature Communications*, *7*(1), 13669. <https://doi.org/10.1038/ncomms13669>
- Craddock, M. (2018). metaSDT: Calculate Type 1 and Type 2 Signal Detection Measures. (R package version 0.5.0). <https://github.com/craddm/metaSDT>
- Dienes, Z., & Seth, A. (2010). Gambling on the unconscious: A comparison of wagering and confidence ratings as measures of awareness in an artificial grammar task. *Consciousness and Cognition*, *19*(2), 674–681. <https://doi.org/10.1016/j.concog.2009.09.009>
- Dotan, D., Pinheiro-Chagas, P., Al Roumi, F., & Dehaene, S. (2019). Track it to crack it: Dissecting processing stages with finger tracking. *Trends in Cognitive Sciences*, *23*(12), 1058–1070. <https://doi.org/10.1016/j.tics.2019.10.002>
- Eichner, C., & Berna, F. (2016). Acceptance and efficacy of metacognitive training (MCT) on positive symptoms and delusions in patients with schizophrenia: A meta-analysis taking into account important moderators. *Schizophrenia Bulletin*, *42*(4), 952–962. <https://doi.org/10.1093/schbul/sbv225>
- Faivre, N., Roger, M., Pereira, M., de Gardelle, V., Vergnaud, J. C., Passerieux, C., & Roux, P. (2021). Confidence in visual motion discrimination is preserved in individuals with schizophrenia. *Journal of Psychiatry & Neuroscience: JPN*, *46*(1), E65–E73. <https://doi.org/10.1503/jpn.200022>
- Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, *34*(10), 906–911. <https://doi.org/10.1037/0003-066X.34.10.906>
- Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 443. <https://doi.org/10.3389/fnhum.2014.00443>
- Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review*, *10*(4), 843–876. <https://doi.org/10.3758/BF03196546>
- Gott, J., Bovy, L., Peters, E., Tziouridou, S., Meo, S., Demirel, Ç., Esfahani, M. J., Oliveira, P. R., Houweling, T., Orticoni, A., Rademaker, A., Boutilik, D., Varatheeswaran, R., van Hooijdonk, C., Chaabou, M., Mangiaruga, A., van den Berge, E., Weber, F. D., Ritter, S., & Dresler, M. (2021). Virtual reality training of lucid dreaming. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1817), 20190697. <https://doi.org/10.1098/rstb.2019.0697>
- Haddara, N., & Rahnev, D. (2019). Trial-by-trial feedback does not improve performance or metacognition in a large-sample perceptual task. *Journal of Vision*, *19*(10), 27. <https://doi.org/10.1167/19.10.27>
- Haddara, N., & Rahnev, D. (2020, March 26). The impact of feedback on perceptual decision making and metacognition: Reduction in bias but no change in sensitivity. *PsyArXiv*. <https://doi.org/10.31234/osf.io/p8zyw>
- Hasson-Ohayon, I., Avidan-Msika, M., Mashiach-Eizenberg, M., Kravetz, S., Rozencwaig, S., Shalev, H., & Lysaker, P. H. (2015). Metacognitive and social cognition approaches to understanding the impact of schizophrenia on social quality of life. *Schizophrenia Research*, *161*(2–3), 386–391. <https://doi.org/10.1016/j.schres.2014.11.008>

- Hoven, M., Lebreton, M., Engelmann, J. B., Denys, D., Luigjes, J., & van Holst, R. J. (2019). Abnormalities of confidence in psychiatry: An overview and future perspectives. *Translational Psychiatry*, 9(1), 268. <https://doi.org/10.1038/s41398-019-0602-7>
- Lange, K., Kühn, S., & Filevich, E. (2015). "Just another tool for online studies" (JATOS): An easy solution for setup and management of web servers supporting online studies. *PLoS ONE*, 10(6), e0130834. <https://doi.org/10.1371/journal.pone.0130834>
- Lebreton, M., Langdon, S., Slieker, M. J., Nooitgedacht, J. S., Goudriaan, A. E., Denys, D., van Holst, R. J., & Luigjes, J. (2018). Two sides of the same coin: Monetary incentives concurrently improve and bias confidence judgments. *Science Advances*, 4(5), eaaq0668. <https://doi.org/10.1126/sciadv.aaq0668>
- Lysaker, P. H., Vohs, J., Minor, K. S., Irrazaval, L., Leonhardt, B., Hamm, J. A., & Dimaggio, G. (2015). Metacognitive deficits in schizophrenia: Presence and associations with psychosocial outcomes. *Journal of Nervous and Mental Disease*, 203(7), 530–536. <https://doi.org/10.1097/NMD.0000000000000323>
- Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, 21(1), 422–430. <https://doi.org/10.1016/j.concog.2011.09.021>
- Massoni, S., Gajdos, T., & Vergnaud, J. C. (2014). Confidence measurement in the light of signal detection theory. *Frontiers in Psychology*, 5, 1455. <https://doi.org/10.3389/fpsyg.2014.01455>
- Morey, R. D., Rouder, J. N., Jamil, T., Urbanek, S., Forner, K., & Ly, A. (2018). BayesFactor: Computation of Bayes factors for common designs (Version 0.9.12-4.1) [Computer software]. Comprehensive R Archive Network. <https://CRAN.R-project.org/package=BayesFactor>
- Nelson, T. O., & Narens, L. (1994). Why investigate metacognition. *Metacognition: Knowing about Knowing*, 13, 1–25.
- Pereira, M., Faivre, N., Iturrate, I., Wirthlin, M., Serafini, L., Martin, S., & Millán, J. D. R. (2020). Disentangling the origins of confidence in speeded perceptual judgments through multimodal imaging. *Proceedings of the National Academy of Sciences of the United States of America*, 117, 8382–8390. <https://doi.org/10.1073/pnas.1918335117>
- Pereira, M., Megevand, P., Tan, M. X., Chang, W., Wang, S., Rezai, A., & Faivre, N. (2021). Evidence accumulation relates to perceptual consciousness and monitoring. *Nature Communications*, 12(1), 3261. <https://doi.org/10.1038/s41467-021-23540-y>
- Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, 117(3), 864–901. <https://doi.org/10.1037/a0019737>
- Rahnev, D., Balsdon, T., Charles, L., de Gardelle, V., Denison, R. N., Desender, K., Faivre, N., Filevich, E., Fleming, S., Jehee, J., Lau, H., Lee, A. L. F., Locke, S. M., Mamassian, P., Odgaard, B., Peters, M. A. K., Reyes, G., Rouault, M., Sackur, J., . . . Zylberberg, A. (2021). Consensus goals for the field of visual metacognition. *PsyArXiv*. <https://doi.org/10.31234/osf.io/z8v5x>
- Recht, S., de Gardelle, V., & Mamassian, P. (2021). Metacognitive blindness in temporal selection during the deployment of spatial attention. *Cognition*, 216, 104864. <https://doi.org/10.1016/j.cognition.2021.104864>
- Rouy, M., de Gardelle, V., Vergnaud, J.-C., Reyes, G., Filevich, E., & Faivre, N. (2021). *Replication: Domain-general enhancements of metacognitive ability through adaptive training*. <https://doi.org/10.17605/OSF.IO/RQ967>
- Rouy, M., Saliou, P., Nalborczyk, L., Pereira, M., Roux, P., & Faivre, N. (2021). Systematic review and meta-analysis of metacognitive abilities in individuals with schizophrenia spectrum disorders. *Neuroscience and Biobehavioral Reviews*, 126, 329–337. <https://doi.org/10.1016/j.neubiorev.2021.03.017>
- RStudio Team. (2020). RStudio: Integrated Development Environment for. <http://www.rstudio.com/>
- Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, 25(1), 128–142. <https://doi.org/10.3758/s13423-017-1230-y>
- Sharp, G. L., Cutler, B. L., & Penrod, S. D. (1988). Performance feedback improves the resolution of confidence judgments. *Organizational Behavior and Human Decision Processes*, 42(3), 271–283. [https://doi.org/10.1016/0749-5978\(88\)90001-5](https://doi.org/10.1016/0749-5978(88)90001-5)
- Shekhar, M., & Rahnev, D. (2021a). The nature of metacognitive inefficiency in perceptual decision making. *Psychological Review*, 128(1), 45–70. <https://doi.org/10.1037/rev0000249>
- Shekhar, M., & Rahnev, D. (2021b). Sources of metacognitive inefficiency. *Trends in Cognitive Sciences*, 25(1), 12–23.
- Singmann, H., Bolker, B., & Westfall, J. (2015). Afex: Analysis of Factorial Experiments (R Package Version 0.15-2). <http://CRAN.R-project.org/package=afex>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., & Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43), 1686. <https://doi.org/10.21105/joss.01686>
- Xue, K., Shekhar, M., & Rahnev, D. (2021). Examining the robustness of the relationship between metacognitive efficiency and metacognitive bias. *Consciousness and Cognition*, 95, 103196.

Received August 5, 2021

Revision received November 21, 2021

Accepted December 14, 2021 ■